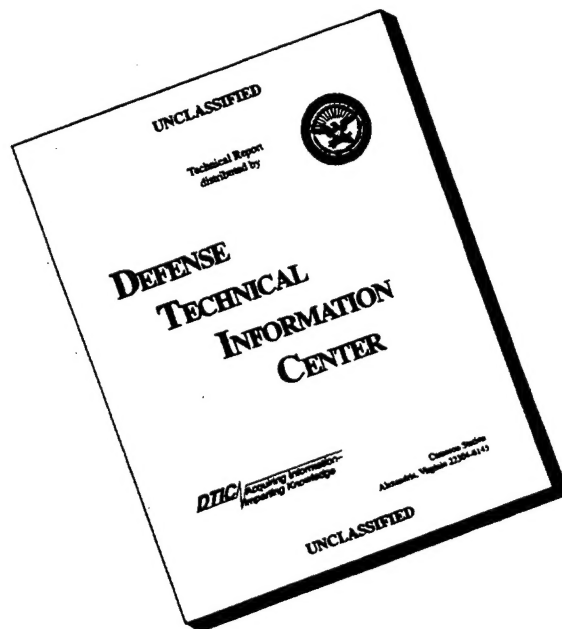


REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE	3. REPORT TYPE AND DATES COVERED FINAL 01 SE091 TO 30 AUG 95	
4. TITLE AND SUBTITLE SOLVING MANUFACTURING PROBLEMS USING SOFTWARE FOR LARGE-SCALE NONLINEAR PROGRAMMING			5. FUNDING NUMBERS F49620-91-C-0079	
6. AUTHOR(S) DR CONN			61102F 7979/00	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) IBM T.J. WASTON RESERCH CENTER P.O. BOX 218 YORKTOWN HEIGHTS NEW YORK 10598			8. PERFORMING ORGANIZATION REPORT NUMBER AFOSR-TR-96	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) AFOSR/NM 110 DUNCAN AVE, SUITE B115 BOLLING AFB, DC 20032-0001			F49620-91-C-0079	
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION STATEMENT APPROVED FOR PUBLIC RELEASE: DISTRUBTION UNLIMITED			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) SEE REPORT FOR ABSTRACT				
14. SUBJECT TERMS			15. NUMBER OF PAGES	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT SAR	

19960520 001

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

Final Report on ARPA Grant Number RF49620-91-C-0079

A. R. Conn

September 29, 1995

Dear Neal Glassman

Below is the final report on ARPA Grant Number RF49620-91-C-0079.

1 PREAMBLE

I would like to begin by stating at the outset how grateful I have been to receive this grant. The grant served as an extremely useful catalyst to pursue applications of optimization to manufacturing problems and I have every intention of continuing this research.

The application of our optimization research to practical manufacturing problems has presented considerable challenges in the current economic environment. Many companies have been involved in the process of restructuring (including IBM), and manufacturing divisions have been hit particularly hard. Some manufacturing facilities have already been closed, while others must either meet ambitious new production targets or themselves face closure. Our work with the manufacturing divisions reflects this uncertainty: one of our projects was terminated by the closure of the facility; a second project was terminated because it was deemed no longer critical to meeting the plant's goals. Moreover, understandably, this is an especially difficult time to be soliciting cooperation from the manufacturing sector. Moreover, even in ideal times it would be unrealistic to assume that we would achieve our ultimate goal (seeing the solution implemented in the operation of the manufacturing site) in more than a few cases.

Nevertheless, we have been working on real problems, collecting real data and have established genuine cooperation with manufacturing personnel. In some cases it is clear that we have had significant success in influencing a manufacturing process and in others we expect similar results, given time.

Moreover, some very substantial research was carried out. In particular, both LANCELOT and CUTE have international reputations and I am very confident that in the future they will have some high profile roles to play in the manufacturing sector. The derivative free work is also very promising and enthusiasm has been expressed by the highest executive levels in research at both the BOEING and IBM corporations.

I certainly hope to apply for ARPA funding again in the future.

2 SUMMARY

Our work under the ARPA grant has proceeded on three fronts: the development of algorithms

to solve large-scale nonlinear optimization problems, the development of methods of optimization that do not require derivatives and are relatively insensitive to noise, and the application of these algorithms to problems in manufacturing.

We achieved our three main objectives in the development of algorithms:

Completion of CUTE (Constrained and Unconstrained Test Environment).

This test environment is a significant contribution to both researchers and practitioners. Researchers can use the provided tools to build interfaces to their own nonlinear optimization software and compare its performance against several other packages on a large database of test problems. Practitioners can state their problems in the standard input format, and then apply several different packages to find the best solution. The standard input format relieves the practitioner of repeatedly inputting each problem according to the specific input requirements of each optimization package. This software has been accessed world-wide by more than two thousand users and for some has become a de facto standard for testing. It has been published in Bongartz *et al.* (1995b) and has already been quite widely cited.

With the completion of CUTE, we have begun the numerical testing required to compare LANCELOT and MINOS, which are the two foremost large-scale nonlinear optimization packages. Once we have completed the generation of the test results, we hope to identify the relative strengths and weaknesses of the packages on broad classes of problems. As with

CUTE, the knowledge gained in this comparative study should be useful to researchers and practitioners alike. Researchers can benefit from our results as they develop new algorithms, and practitioners will know which software package is best suited to solving their problems.

Testing of LANCELOT

We have completed intensive numerical testing of LANCELOT on more than nine hundred test problems. These tests helped in identifying some enhancements to the algorithm, and also form the basis for our comparison of LANCELOT with MINOS in Bongartz *et al.* (1995a). Thus, for example, our results where the number of inequalities greatly exceeded the number of original variables motivated our analysis in Conn *et al.* (1994b). The testing also resulted in our investigation of penalty parameters for each constraint and motivated a separate treatment for linear constraints in Conn *et al.* (1993a) and Conn *et al.* (1993b) that is to be published in Conn *et al.* (1995c).

Derivative free optimization

In the application area there is a real need for methods of optimization that do not require derivatives. For example, so-called pattern search methods such as Nelder and Mead (1965) are still popular with practitioners although significantly better methods exist. Moreover, because of the nature of the applications, such methods should be able to handle noise. Motivated by some work of Powell (see Powell, 1994a and Powell, 1994b), we explored the use

of multivariate interpolation techniques in the context of such methods for unconstrained optimization. We used low order polynomial models in a trust region framework. A prototype algorithm exists in MATLAB 4.1.1 (see The Mathworks Inc, 1992) and is being used in a design of experiments environment to attempt to optimize the result of a complex and costly numerical procedure required in the analysis of the in flight vibration of a helicopter rotor. This is to appear in Conn and Toint (1995).

In applying nonlinear optimization algorithms to problems in manufacturing, we have tried to find practical problems faced by manufacturing engineers that are nonlinear and preferably of generic interest. Our objective has been to solve problems that cannot be solved adequately by existing methods. We have had some success throughout the duration of this grant, and we are hopeful that the work on some of our current problems will make useful contributions to manufacturing. Below we list the applications on which we have worked, together with the manufacturing site:

- batch sizing and scheduling the assembly and test of hard disks (IBM Rochester)
- maximizing the throughput of the photolithography sector of a semiconductor manufacturing line (IBM Burlington)
- identifying the region of feasible solutions in manufacturing design problems (Dassault France)
- minimizing the maximum delay in small transistor circuits (IBM Kingston)
- tuning circuits (IBM Austin)
- quality control and testing of components on computer boards (IBM Poughkeepsie)
- Optimization of the design of complex systems in the aeronautical industry (Boeing Corporation)

In addition to the applications above, we have developed algorithms to solve several difficult classical applications problems in optimization:

- location-allocation and related problems
- macro-cell placement problems
- discontinuous problems, including fixed charge problems, nonlinear robust regression, cutting stock problems, etc.
- problems with noise and problems where derivatives are unavailable

These algorithms offer significant improvements over existing methods.

The algorithmic and applications work combined have yielded twenty publications that have benefitted from and acknowledge ARPA support.

We plan to continue work on most of the applications listed above, with a few exceptions. In particular, the focus of the work with IBM Burlington has shifted from optimizing the through-

put of just the photolithography sector to modeling the entire semiconductor line with a view to reducing cycle time. Their current priorities do not encompass optimization so this project is unlikely to continue in the near future. In addition, the IBM Rochester ADSTAR manufacturing facility has been closed, and so work on the batch sizing and scheduling problem is currently stopped. If possible, we will continue this research with a similar manufacturing facility in IBM

San Jose. More importantly, we continue to actively develop real applications in the manufacturing sector.

We now give more details on the research outlined above.

3 Development of Algorithms

3.1 CUTE (Constrained and Unconstrained Testing Environment)

The problem of inputting nonlinear problems in thousands of variables, maintaining a database of problems and being able to test different packages, all of which expect their own particular form of input, is an important and nontrivial task. We have developed tools to probe the data set, aid in the evaluation of results and to assist in interfacing the standard input format for LANCELOT to other packages. These tools have been written as production software that can be installed automatically and are currently available in the public domain. In particular, we have an interface that will take any problem input intended for LANCELOT and preprocess it into a suitable format for use with MINOS. Since MINOS and LANCELOT are the two most suitable production software packages available for large-scale nonlinear optimization, this interface will be very useful to many researchers. In addition we have interfaces to several HARWELL subroutines for optimization, the unconstrained and tensor minimizers of Bobby Schnabel, and IBM's OSI. Since its release, our software has been ftp-ed worldwide. The set of test problems is undoubtedly the most extensive database of nonlinear programming problems and the ability to readily test alternative optimization software on problems written in this format is likely to be invaluable for algorithmic development in the field.

Following its release, we have worked on installation scripts for new platforms and interfaces for additional optimization packages, including an interface to NPSOL and interior point techniques. Moreover we have interface tools with MATLAB, (The Mathworks Inc, 1992). Since MATLAB is frequently used for prototyping algorithms, this should be especially useful for researchers in optimization.

This research was recently published in Bongartz *et al.* (1995b)

3.2 Testing and Enhancement of LANCELOT

Much of our work in this area has been based upon learning from our extensive numerical experience to date. Thus we have recognized the importance of accounting for the special structure of slack variables and now have an adequate approach that motivates a full implementation. The details are given in Conn *et al.* (1994b).

We have considered the implications of various formulations of a particular group partial separable structure and have been able to show how relatively simple criteria can be remarkably effective at improving a decomposition. This work was published in Conn *et al.* (1994a). Moreover, we have been able to generalize the idea of a single trust region to exploit the structure inherent in partial separability. This work appears in Conn *et al.* (1995d).

In the context of interior point methods it is not clear how to exploit a suitable trust region model and thus we need to investigate both this and line search approaches. Much of our

discontent with the augmented Lagrangian framework has been the result of our numerical experience with linear constraints and this has led us to consider analogous methods for interior point techniques in addition to the treatment of linear constraints separately. This resulted in the publications Conn *et al.* (1995e), Conn *et al.* (1995f), Conn *et al.* (1994c), and Conn *et al.* (1995a).

We are also using our numerical tests to accumulate numerical data concerning the efficacy of various preconditioner updating techniques and the handling of degeneracy, for example. Besides the ongoing work (see, Bongartz *et al.*, 1995a and Conn and Toint, 1994 — these projects tend to be somewhat long term) the results concerning LANCELOT alone have been published in Conn *et al.* (1988) and Conn *et al.* (1995b).

3.3 Comparison of LANCELOT and MINOS

For general large-scale nonlinear optimization problems, only two production level software packages are readily available: LANCELOT and MINOS. Considering that MINOS has been around since the 70's and is well-known, it is amazing that, with the exception of linear programming results (for which it was not explicitly designed), its numerical performance has not been extensively reported in the literature. The linear programming performance was reported when the initial comparisons with Karmarkar's interior point approach were made, presumably because of the lack of access to other commercial linear programming packages. We can only assume that the results with large-scale nonlinear problems were not reported because of the lack of other available codes and because of the amount of effort required to create a suitable environment for such testing. With the completion of CUTE, we have begun comparing the performance of these two packages on a set of test problems that suitably characterize the available nonlinear programming database. Since some of the test problems are quite large (more than 15,000 variables), generating all the test results is an extensive task. From these results, we hope to identify the relative strengths and weaknesses of these packages on broad classes of problems (e.g., primarily linear problems, highly nonlinear problems, etc.). Once again, we should not underestimate the general level of usefulness of results of this type. We are fortunate in that we are able to produce a report jointly with one of the authors of MINOS (Michael Saunders).

3.4 Derivative free Optimization

We consider the problem of minimizing an objective function whose value is determined by measuring some quantity in the real world. This measure may be of a physical nature (for instance, the depth of a certain layer in geophysical exploration) or be related to other contexts. We focus on the case where there are no constraints on the problem variables although the generalization to simple bounds is quite straightforward. Moreover, since the proposed method is derivative free, one might want to handle constrained problems using an exact penalty function.

By extension, we also consider unconstrained optimization problems whose objective function is the result of a complex and costly numerical procedure, possibly involving some considerable noise (due, for instance, to truncation or approximation in the calculation defining the objective).

We explore the use of multivariate interpolation techniques in the context of methods for unconstrained optimization that do not require derivatives of the objective function. A new algorithm is proposed that uses quadratic models in a trust region framework. The algorithm is constructed to require few evaluations of the objective function and is designed to be relatively insensitive to noise in the objective function values. The details, including numerical results, are published in Conn and Toint (1995).

3.5 Iterated Subspace Minimization

The most common methods for unconstrained minimization either determine a search direction followed by a linesearch or use the trust-region approach (see, for example, Dennis and Schnabel, 1983). In the former case, a simple model of the underlying objective function is constructed in order to determine the search direction. By contrast, in the latter case, an approximate minimizer of the model within a restricted domain (the trust region) is determined. This model minimizer is then used as a prediction of the actual minimizer of the true objective. In a trust-region method, success of this process is measured by comparing the model and true function values at the predicted minimizer. In linesearch methods, the true function is used to establish a step size. Thus, both of these approaches may be considered to perform their multi-dimensional work with respect to a model whilst probing the true function uni-dimensionally. Of course, the model does make use of the true function and perhaps its derivatives — maybe at more than a single point.

In this paper, we take the view that the above schemes are quite wasteful, given the amount of information that may have been accrued during the (approximate) minimization of the model. In particular, the model may have been sampled in a number of potentially interesting directions, of which only the aggregate direction is normally considered to be of significance.

We also believe that, provided function and derivative values are inexpensive to compute relative to the linear-algebra costs, an (approximate) low-dimensional minimization is a relatively simple calculation. Indeed, we feel that there is high-quality, robust, general purpose software readily available for the small scale unconstrained minimization problem, and that such software is normally capable of solving problems of modest dimensions - say up to 100 variable problems - extremely fast on current workstations *provided that* function evaluation is cheap. Of course there are, and will continue to be, small-scale problems which are challenging, because they are so nonlinear that algorithms implemented in fixed, finite precision arithmetic are unsuccessful, but in our experience such examples occur rarely in practice.

Thus, in this paper, we propose methods which aim to investigate the *true* objective function in a space larger than the one-dimensional space which is normally associated with linesearch or trust-region methods. We do this knowing that, so long as the space is relatively modest, the approximate multi-dimensional minimization will still be a manageable calculation. Moreover, by carefully choosing the space that we investigate, we hope to reduce significantly the linear-algebra costs while still maintaining global, and fast asymptotic, convergence.

1 Application of Algorithms to Problems in Manufacturing

4.1 Batch-Sizing and Scheduling

The Rochester ADSTAR manufacturing facility produced files (hard disks) for computers ranging from micros to minis. These files were manufactured not only for IBM products, including PS/x's, RS/6000's and AS/400's, but also for a growing OEM market. The components required to manufacture the files were provided by outside suppliers. The files belonged to three families, where some families contained several models; models were differentiated by capacity, density, and performance.

The two main functions in the manufacturing process were assembly and test. In the assembly area, the files passed through three stages, where each stage had several lines and all lines accommodated all file types. A set-up time was required to switch a line from one file type (or sometimes, from one file family) to another. At demand levels, the manufacturing line was operating close to, and in some cases beyond, capacity. The testing area in particular was having difficulty meeting the demand for certain file types, thereby creating backlogs of files waiting to be tested. In addition to these capacity shortages, there were parts shortages for some components, such as the disks and actuators. Thus the manufacturing and component resources had to be carefully managed to ensure Rochester met its weekly demands for files, or at least satisfied as much of this demand as possible.

In cooperation with Rochester, we were constructing an optimization model to solve this file production problem. Ultimately, the model should have solved both the batch-sizing and scheduling problem. The batch-sizing problem would have decided how many of each file to produce each day, while the scheduling problem would have determined in what order the batches should have been produced. The batch sizes were to be chosen to meet as much of the demand as possible, subject to constraints on production capacity and component availabilities.

We completed a batch-sizing model of the assembly area which decided not only the quantity of each file type to produce each day, but also allocated this production to particular lines within the assembly area. We also built a preliminary model of the testing area to assess the usage of the test facilities. With the assembly and test area data supplied by Rochester, we generated some preliminary results. The batch-sizing model as solved involved approximately 2500 variables and 2000 constraints, not including simple bound constraints. The optimization enabled us to analyze several potential scenarios and in particular the bottlenecks were clearly indicated. We prepared an internal technical report describing these details. Because of the production data, this internal report is confidential, but we plan to prepare a report for the public domain that describes the solution methods but does not include confidential data.

The following table gives some sample output from the batch-sizing model for the assembly area, under the assumption that each day we have only enough inventory of each component to produce one-fifth of the weekly demand.

(To avoid releasing confidential data, we have altered the numbers in this and the subsequent table, but the qualitative results are genuine.)

In the scenario above, the output of the assembly area is constrained by the limited component

Model	Day 1	Day 2	Day 3	Day 4	Day 5	Total
1	99	121	110	110	110	550
2	771	771	771	771	772	3856
3	376	376	376	376	375	1879
4	1257	1257	1257	1257	1257	6285
5	266	266	266	266	267	1331
6	486	486	486	486	486	2430
7	183	179	184	186	183	915
Daily Production	3438	3456	3450	3452	3450	17246

inventory. If we assume that we have enough inventory available each day that we could produce the entire weekly demand in one day, we get the following results:

Model	Day 1	Day 2	Day 3	Day 4	Day 5	Total
1	0	0	0	0	564	564
2	1780	618	628	443	505	3983
3	0	187	469	702	521	1879
4	2817	0	627	968	1873	6285
5	0	0	302	611	418	1331
6	338	1091	1099	551	0	3082
7	202	460	0	0	253	915
Daily Production	5146	2356	3125	3278	4134	18039

Under this scenario, the output of the assembly area increases by 4.59%. Now, instead of being constrained by the limited component inventory, we are constrained by the available production time on one of the assembly lines. This is just one example of the what-if scenarios that could be run on the batch-sizing model.

Unfortunately, at this point, the Rochester manufacturing facility was closed and work on this problem is currently stopped. Rochester's ADSTAR manufacturing operations have been consolidated with a similar facility in IBM San Jose, and we hope to continue our research with this site.

The model as first proposed to Rochester assumed that production was scheduled on a build-to-order basis and that capacity would not always be sufficient to satisfy demand. In this environment, the objective would have been the maximization of a nonlinear priority function, where this function reflected the demands and due dates of orders for the various files. Later, for the purposes of the initial model at least, we were informed that production was actually scheduled on a build-to-plan basis, with relatively constant quantities of each file to be built each week. The objective of maximizing the nonlinear priority function was therefore replaced with the objective of maximizing total production, which is a linear function. The resulting batch-sizing

model under this objective is a large mixed integer (i.e., there are both continuous and integer variables) programming problem (i.e., the objective function and the constraints are all linear). Thus, in its current form, the batch-sizing model does not require large-scale nonlinear programming techniques. It could instead be solved by software designed for large-scale mixed integer problems. Nevertheless, the flexibility of our model allows us to easily return to the objective of priority maximization, if it turns out that this is in fact the better approach.

4.2 Stepper Throughput Problem in Photolithography

Photolithography is the process by which a microscopic pattern is transferred from a photo mask to a material layer in a circuit. For example, one might etch into a silicon dioxide layer on the surface of a silicon wafer.

A simplified version of the wafer manufacturing process is as follows: the oxidized wafer is first coated with a layer of a light sensitive material called photo resist and then exposed to ultraviolet light through the photo mask. The exposure renders the photo resist insoluble in a developer solution; hence a pattern of the photo resist is left wherever the mask is opaque. The wafer is next immersed in an acid solution, which selectively attacks the silicon dioxide, leaving the photo resist pattern and the silicon substrate unaffected. In addition, MOS circuit elements typically require negative and/or positive dopants (for example, boron and phosphorus) that are implanted via ionization. In the final step, the photo resist pattern is removed by means of another chemical treatment.

The cameras for this process, referred to as steppers, are very expensive and the control of the lens and blades is an inherently slow process. A typical sequence of operations on the stepper involves loading a computer program, loading a reticle (mask for one or perhaps several chips), setting the blades (to protect the rest of the wafer), loading the wafer, doing a rough alignment, refining this alignment, and then stepping (i.e., moving from chip to chip) and exposing, with many loops involving resetting of blades, realignment, etc. Several steppers are available on the process line but not all chips can be run on all models. Furthermore, demand often exceeds capacity. Various options (increasing the number of operators, buying new masks which expose several chips simultaneously, reducing the number of alignment sites, etc.) are available to increase the stepper throughput, but these options can be costly. The problem is to identify those changes which yield the greatest improvement in throughput at the least cost.

We built two versions of the stepper throughput model, namely one to minimize production cost subject to constraints on production time and demand, and one to minimize production time subject to constraints on production costs and demand. The model was based upon empirical data and attempted to determine the best investments to make to improve throughput. The original model included only one tool set, namely the Nikon steppers, and did not allow the possibility of buying new masks to expose several chips simultaneously. The model was first extended to include

the option of purchasing new masks, and later extended to include a representation of another tool set. With the additional tool set, we aimed to determine how production of the wafers should be distributed over the two sets to increase throughput. However, Burlington recently decided that stepper throughput was no longer their biggest problem and their efforts should be directed

towards developing a model of the entire production line, rather than a single gate in this line.

Work on the stepper throughput model has therefore stopped and interest has shifted to the more global situation instead. We refer to this new model as the 'bubble phenomenon' and give details in the next section.

Before we began to work with Burlington, they had developed a nonlinear function which calculated the throughput of the Nikon steppers based on the number of alignment sites, the number of blade settings, the number of exposures per wafer, etc., but they did not know how to maximize this function, especially if constraints were imposed on the allowable ranges for some variables. To 'analyze' this function, they set all but one of the variables to their current values, and then evaluated the throughput function for several different values of the remaining variable. This approach told Burlington what the first-order effects of each of the variables were, but could not tell them what the best overall solution would be. We were not only able to maximize their original throughput function, but we were also able to include additional terms for new production variables (such as purchasing new masks) and to include constraints on the production variables. It should be noted that the throughput function is inherently nonlinear, and therefore linear programming techniques would not be appropriate for this application.

We will use an example to illustrate the results provided by our model. Suppose our objective is to minimize daily production cost subject to satisfying demand. The following table shows the best changes to make to improve stepper throughput as the daily demand is increased:

Daily Demand	Change to Improve Stepper Throughput
1000 wafers/day	Purchase new masks to reduce SPW (number of shots per wafer) to 30.
1100	Purchase new masks to reduce SPW to 25 (the minimum possible) and hire additional operators to act as work coordinators.
1200	Purchase new masks to reduce SPW to 25 and hire as many additional operators as possible.
1300	Demand can no longer be satisfied.

(To avoid releasing confidential data, we have altered the numbers, but the qualitative results are genuine.)

The results above assume all production is run on the Nikon stepper. If each wafer requires 15 passes and at most five of these passes can be run on the Perkin Elmers (another tool set), we get the following results:

Daily Demand	Change to Improve Stepper Throughput
1400 wafers/day	Run 11 passes on Nikons, 4 passes on Perkin-Elmers. Also purchase new masks to reduce SPW to 25.
1500	Run 10 passes on Nikons, 5 passes on Perkin-Elmers. Purchase new masks to reduce SPW to 25 and hire additional operators.
1600	Run 10 passes on Nikons, 5 passes on Perkin-Elmers. Purchase new masks to reduce SPW to 25 and hire as many additional operators as possible.
1700	Run 10 passes on Nikons, 5 passes on Perkin-Elmers. Purchase new masks to reduce SPW to 25 and hire as many additional operators as possible.
1800	Reduce the number of alignment sites. Demand can no longer be satisfied.

4.3 Bubble Phenomenon

The Burlington semiconductor manufacturing line in Building 973 manufactures 5-inch logic wafers. The engineering staff which runs this line is currently faced with the challenge of reducing this line's cycle time by 30%. In the course of studying the flow of wafers through their line, the engineers have discovered that this flow is not always constant, but sometimes contains significant 'bubbles', which consist of many wafers at one gate in the line. These bubbles are created when one gate falls behind on its production, usually because of equipment breakdowns at that gate. The gates most likely to cause these bubbles are those that have only a few tools. When all the tools in such a gate are operational, the capacity significantly exceeds the production requirement. With even one breakdown, however, the capacity of the gate falls significantly below required levels. For example, an ion implant gate with three tools may run at 120% of the daily going rate when all three tools are operational, but this level drops to 80% when one tool goes down. The wafers then accumulate in front of the ion implanters, waiting to be processed. During this time, subsequent gates might be starved for parts, and they too will be idle. Once all the ion implanters are operational again, the wafers will be processed at a rate that exceeds the capacity of subsequent gates in the line. The wafers will again accumulate in front of these subsequent gates, and the bubble will persist in the production line for many days.

Burlington would like to model the bubble phenomenon, and use the model to identify changes in the line that would help the wafers flow more smoothly through the line. Reducing the WIP (work in process) should lead to a reduction in the line's cycle time.

To begin, we began by modeling not the full 5-inch semiconductor manufacturing line, but rather a mini line which consists of fourteen gates. These gates were chosen by Burlington to be representative of the most important gates in the full line. We formulated an initial model and

Burlington now need to gather the necessary test data.

Burlington already has a detailed simulation model of their semiconductor manufacturing line. This simulation, however, does not run much faster than real time. (We do not know the ratio of simulation time to actual production time, but it would certainly take many hours to model one month of production.) To address the bubble phenomenon with the simulation tool, an informed user would have to run a simulation based on the current production line, identify changes that might improve the bubble phenomenon, run a simulation on the changes, use the simulation output to identify new changes, and so on, until a set of changes had been identified that would alleviate the bubble phenomenon. Even then, the identified changes would not likely be the best possible ones, but rather only the best of those simulated. (What 'best' means in this application must be decided by Burlington. It might, for example, mean the set of changes which minimizes WIP subject to budget constraints.) With the current simulation tool, this process could easily take weeks. We expect that our optimization model will run significantly faster and identify the best improvements to the line. Unfortunately, however, Burlington's current priorities have prevented them from continuing on this project at this time.

4.4 Manufacturing Design Problems

Dassault France have a generic set of manufacturing design problems. Typically their designs (for an entire aircraft body, for example) involve a number of variables, inequality constraints and equations. The number of variables can be large, can include some logical and integer variables and at least some of the relationships can be highly nonlinear. What they desire is to be able to obtain some general idea of the entire region of feasible solutions so that in the manufacturing process, they have the possibility of choosing components that are preferable on the basis of other (subjective and objective) criteria.

To date we have demonstrated that we are able to develop such solutions, and indeed, in the case of the design of an aircraft structure, we have been able to indicate to them that they have not been solving the correct problem, because of inconsistency in the data.

We are currently modeling a complex engine gearing system for which, according to Dassault, no solution has ever been determined.

Eventually they would like to incorporate our techniques into a generic design tool, if this is possible.

It is hard to be more definite at this time as to the eventual benefits since we are still at a relatively early stage. However, from the discussions and experiences that I have had with Dassault, France it would appear that we are solving the problems much faster than they expected, which is very important from the point of view of incorporating our techniques into a generic design tool.

It should be stressed that these problems are inherently very nonlinear and it is essential that nonlinear techniques be used.

On another front, we are applying the derivative free optimization algorithm to optimize complex systems using both design of experiments models and methods related to pattern search techniques. This is joint work with Rice University and the Boeing Corporation.

Typical applications generate very expensive function evaluations, often determined via sophisticated simulation packages. Consequently the 'outer model' provided by the simulation does not readily lend itself to straightforward nonlinear optimization. For example in a quasi-Newton sequential quadratic programming context the cost of determining and solving successive quadratic programs would be prohibitive, even assuming first derivatives are available from the simulation. However, these types of engineering optimization problems arise frequently.

In the context of this joint project, the particular problem of interest is optimization of the engineering design of a helicopter rotor blade so as to minimize vibrations for specified flight conditions. This design is sufficiently complex and expensive so as to make the use of mathematical models imperative. There is a solid basis of computational fluid dynamics, structural analysis and control theory that theoretically allows standard nonlinear optimization to take place but although such a system is deterministic, realistic optimization for even moderate size problems is prohibitively expensive and slow.

Consequently, the main thrust of the project is to explore approximate models over which it is more reasonable to optimize. The immediate questions that arise are

- What constitutes a suitable approximation?
- What is an appropriate optimization technique?
- What represents a 'solution'?

Moreover these basic questions are clearly related in that initially, when far from the solution, it would be efficient to have a relatively crude model, optimization technique and stopping criterion. However, ultimately, when in the neighborhood of a solution, one could expect to be more careful. Finally, given the relative accuracy of the modeling technique, the presence of noise and perhaps, the absence of derivatives, it makes no sense to use a stopping condition that aims to be more accurate than, for example, the noise threshold.

At Boeing, Paul Frank has been working in the area of optimization of approximate models over the past year and it is clear that Boeing intends to continue with this project. For some time John Dennis and Virginia Torczon have been working productively with the multidisciplinary optimization (MDO) group at Boeing Computer Services. Moreover, my colleagues Katya Mints (a doctoral student at Columbia University who spent the summer of 1995 at T. J. Watson Research Center) and Philippe Toint (University of Namur), and I are working on an extension of the algorithm cited above and published in Conn and Toint (1995). Thus this collaboration involves a strong commitment accompanied by matching and complementary skills and we expect it to be an ongoing project with significant applications.

4.5 Circuit Design

This encompasses two applications. The first concerns the design of small circuits and was carried out through Philip Strensky of the T.J. Watson Research Center.

We are concerned with the design of small circuits (less than 100 transistors) with few inputs and few outputs. The objective is to make the circuit very fast. In our initial application we used a simulator to run some standard patterns on the circuit and measured the delay, which is a function of the transistor widths. Our objective is to minimize the maximum delay as a function of the widths, subject to simple bound constraints on the widths. We formulated this as a nonlinear programming problem. We used about twenty widths, although this could increase. The expense in solving the optimization problem is in the simulation (equivalent to a function evaluation). Thus we fitted a quadratic approximation to the delay function. This gave us a local approximate surface and we then iterated on the approximations, using LANCELOT. The results were used with success in the IBM Kingston plant. We would like to consider using the simulator directly in the optimization (without fitting a quadratic approximation), but we then need to account for noise.

In general, we feel there is much scope for synthesizing optimization techniques with simulation systems and we would like to pursue this further. There has been some interest shown by Burlington Technology Simulation, who use a sophisticated simulation system in the design of semiconductor wafers. This is like a much bigger version of the Kingston circuit problem. However, we have yet to meet with this group at Burlington.

The algorithm is being used at two sites in the USA (Kingston and Austin) and one site in Germany (Boeblingen).

Previously "optimization" was done by informed users making intelligent guesses, doing simulation and then moving around in the design space. The use of more sophisticated nonlinear optimization techniques has given solutions that are 10% better — which in practice is a great deal.

More recently we have considered a direct extension of this work which includes analytic expressions and does not use simulation. It turns out that the optimization in this case can be more efficiently carried out using MINOS and as a direct consequence of this work Dr. Strensky is pursuing the idea of obtaining suitable licenses for MINOS.

Once again, I hope that this will be an ongoing project.

The second application was carried out with Chandu Visweswariah and Ruud Haring of the T.J. Watson Research Center and involves the tuning of circuits. Many circuit designers spend a great deal of time manually sizing their schematics to meet delay, power and area targets. Circuit simulation is often carried out in the inner loop of this iterative and "optimization" is typically a manual process. Instead we have incorporated a modified version of LANCELOT to replace the manual process with real optimization.

The results to date with a prototype research version have been very encouraging and we hope that this will be a continuing and very successful application. Should that be the case there is much work that remains to be done to ensure robustness and adequate starting and stopping criteria in the presence of noise. Moreover we expect to obtain useful feedback from circuit designers who will be trying to use LANCELOT to design real and large circuits.

4.6 Quality Control and Testing

This problem concerns the testing of boards on mainframe machines with a view to obtaining the exact cause of failure so that the particular supplier of the defective component can be identified.

Suppose one has a board, containing many components, that is failing. Typically one replaces a subset of the components with new components to correct the failure. Since the components come from several different manufacturers, we would like to be able to identify the source of the malfunctioning component. Due to cost and diagnostic constraints, however, the exact cause of the failure might be unknown and we are necessarily restricted to incomplete observations for much of the data. However, it is essential that we have some complete information as a result of subsequent testing. Confronted with this real situation we would like to exploit all the available information in an optimal way.

The optimization function is thus to maximize a likelihood function subject to simple bounds on the variables. The problem is difficult in that the growth in the number of variables (because of the large number of conditional probabilities involved) is rapid as the number of components in the subset increases.

To date we have optimized a simple instance and we have explored a more general instance with equally encouraging results. With B. Reiser of the University of Haifa, Israel, Betty Schultz and myself of our group here at Watson we have collected our results in Reiser *et al.* (1994). We would like to generalize this framework in order to obtain confidence intervals using simulation. The method of choice, suggested by the statisticians, is to use the bootstrap technique, Efron and Tibshirani (1986), for obtaining confidence intervals. This would require repeated optimization using computer simulated data. Depending upon the direction taken, again determined by the expertise of the statisticians, there is a possibility that this work will continue. Eventually, it is hoped that the approach will have wide application to quality control and reliability testing within manufacturing.

The current standard practice, when, for example, a board containing k components fails with a probability p , is to assign equal probabilities (i.e., p/k) to the possibility that any given component fails. The proposed approach is inherently superior and moreover it is evident from our current experience that we are able to obtain useful solutions to problems that are too difficult to analyze otherwise.

4.7 Location-Allocation

Location-allocation problems are concerned with the optimal service or supply of a set of existing facilities whose locations are fixed and known. Service of the existing facilities is achieved by a set of new facilities, where the location of each new facility must be determined and its output allocated to some particular subset of the existing facilities. The locations and allocations of the new facilities are chosen to minimize some cost function.

Many problems can be characterized as location-allocation problems, including the location of warehouses or factories to serve customers; the location of public facilities such as ambulance depots, fire stations, or schools to serve surrounding neighborhoods; the optimal layout of machines in a plant, or the addition of new machines to existing plant layouts; the partitioning into

subgroups of circuit modules in VLSI layout; and the problem of cluster analysis, where the data points are treated as existing facilities, and the cluster centroids as new facilities.

We have completed our basic algorithmic work on the uncapacitated algorithm. We performed extensive numerical tests to compare our algorithm with other competitive algorithms. These tests revealed several areas where our algorithm's efficiency could be improved, and these improvements were incorporated into our current software. The resulting algorithm is not faster than alternating methods, although its solution times are comparable with alternating methods and faster than all other methods. More importantly, however, it consistently gives better quality solutions than all other methods and is significantly more general in its approach. This work has been published in Bongartz *et al.* (1994).

This additional flexibility has enabled us to consider the generalization to the capacitated problem. We have completed the theoretical work to extend our approach to include capacity and demand constraints and we have written and tested a preliminary version of the code in MATLAB. We are currently looking at ways to improve the code (starting points, degeneracy handling, etc.)

Our previous work on location-allocation problems enabled one of us (Ingrid Bongartz) to contribute to the development of a "logistics planning tool" for IBM's spare parts distribution system in the United States. This work was done jointly with Management Technologies of the IBM Consulting Group in Dallas, TX, and was not funded through the ARPA grant. The client and funder was IBM U.S. Logistics in Mechanicsburg, PA.

The problem addressed by the logistics planning tool is more general than the location-allocation problems considered previously. In the logistics planning problem, the facilities to be located are distribution centers, in which the spare parts are stored and from which they are shipped to customers whose equipment has failed. The tool determines not only the location of the distribution centers and the assignments of customers to these centers, but the inventory levels for each part at each distribution center and the mode of transportation to be used to transport parts to each customer. The demands are based on the expected failure rates of the parts. The system must be designed to satisfy stringent service constraints, which stipulate the percentage of replacement parts which must be received by the customer within certain elapsed times (e.g., 4 hours or 24 hours) from failure diagnosis. The objective is to minimize the overall costs of the distribution system, including the costs of operating the distribution centers, owning and handling the inventory, and transporting the parts to customers.

Although the logistics planning tool does not use any of the algorithms we developed for location-allocation problems, our expertise allowed us to quickly understand the modelling requirements and to improve the initial algorithms developed by Management Technologies. The solutions recommended by the logistics planning tool have the potential to save IBM U.S. Logistics millions of dollars. This work reaffirmed our belief that location-allocation problems have important practical applications and are therefore a worthwhile topic of research.

In a problem related to the location-allocation problem, Michael Overton of the Courant Institute, New York University (whilst visiting Watson during the summer) and I extended our earlier work on minimizing the sum of Euclidean norms. We define a primal-dual interior-point

method that is easy to motivate and implement. Furthermore, it is very robust and converges very rapidly even when many norms in the summand are zero. The key observation which makes the primal-dual algorithm possible is that Newton's method should be applied to a certain

complementarity condition. The method is apparently as robust as primal-dual interior-point methods for linear programming, in the sense that almost any choice of steplength rule and almost any rule for reducing the centering parameter will usually work, including, in most but not all cases, just setting the steplength to one and the centering parameter to zero. A prototype MATLAB code has been used to compute optimal Steiner trees given their topology and should be useful in many other applications also. In particular, although we originally intended to just publish an internal report, there has been enough external interest in the results that we are now motivated to write a full paper. Two applications that we are aware of that have made direct use of our results are in plasticity, (Anderson and Christiansen, 1995) and image reconstruction, (Chan *et al.*, 1995).

4.8 Vision Sensor Planning Problem

The vision sensor planning problem arises in robotic vision tasks. It refers to the calculation of parameter values for vision sensors that allow automated machines to perform predetermined functions. Steven Abrams and Peter Allen, of Columbia University, and Konstantinos Tarabanis, of Manufacturing Research at IBM Thomas J. Watson Research Center, have formulated the sensor planning problem as a nonlinear optimization problem. This problem consists of determining the sensor's lens center, viewpoint and focal setting in order to maximize the sum of several different feature detectability criteria. These criteria are also the constraints of the optimization problem. They ensure that all features on the object to be viewed are visible (i.e., not occluded), are in focus, are within the sensor's field of view, and appear large enough on the image plane to be seen by the vision system.

At the outset we worked together with Abrams and Tarabanis to state this problem in SIF format and then solve it using LANCELOT. Because some of the constraints are nonsmooth, LANCELOT had difficulty in converging to optimal solutions. Unfortunately this work did not progress beyond the initial numerical experiments although we had some interesting idea for handling the nonsmooth nature of the problem. The group in Manufacturing Research that was investigating sensor planning problems was sold by IBM.

4.9 Layout Problems

The aim of this work is to develop a projected penalty function technique to solve the macro-cell placement problem, implement the results in a working algorithm and compare the algorithm with methods currently being employed. It constitutes joint work with David Mates (MacDonald, Dettwiler and Associates Ltd) and Anthony Vannelli (University of Waterloo) and is the major body of the doctoral thesis of the former.

The macro-cell placement problem is the problem of placing irregularly shaped cells (in our case we consider only rectangular shapes) on a surface such that the distance between those cells that are connected (specified before hand) is minimized. There are many important applications

of this problem in engineering design. Some examples are printed circuit board design, very large-scale integration, operations research location-allocation problems and plant layout problems. This is the first time that this problem has been modeled exactly and solved with a piecewise linear mathematical program.

To accompany the projection algorithm, a computer graphics package was developed to display the current layout for any given iteration. This package makes interpretation of the results much easier as well as allowing the user to manipulate the layout to aid in the choice of starting points or to restart the problem in the hopes of moving out of a local minimum. The package includes an optional eigenvector-based starting point method of Rowan and Vannelli. Together with the projection algorithm code, the graphics package provides a significant step towards creating a state-of-the-art placement package.

Although the solutions are very good, the performance of the code itself is significantly less favorable. The current implementation of the algorithm is undesirably memory intensive and extremely slow. This is largely due to the nature of the algorithm; solving 1000 least squares problems for matrices as large as 5000 by 400 can be time consuming (not to mention the other steps taken in every iteration). However, there are some improvements that should be made, such as exploiting sparsity and reducing the problem size by considering only a subset of the possible constraints.

Performance notwithstanding, the overall results of the package from the test cases considered to date are very encouraging and merit further research of the optimization technique for problems of these types as well as more work on the package itself.

Current methods for the macro-cell placement problem are based on simulated annealing and the Tabu method. In many cases, the problem is solved by hand using computer-aided design tools to move cells on a display screen. This is the first time that a projection method of an exact penalty function has been used. For every problem tested, this new approach produced better quality solutions. However, the comparisons were done on standard cell problems since no macro-cell packages were available at the time the testing was done. Our current preliminary package does not make use of large sparse solvers and so is unnecessarily computationally expensive. However, as an indication of possible improvements, we implemented a version that computed the search direction using a sparse solver. Memory requirements went down from 83 M Bytes of RAM to 19 M Bytes and the new version was over thirty times faster.

This work has been published in the thesis of David Mates at the University of Waterloo, which I and Dr. Vannelli directed (see Mates, 1993).

4.10 Discontinuous Problems

This research involves applying piecewise differentiable techniques to discontinuous problems and was the main body of research carried out by my former doctorate student Marcel Mongeau, now a faculty member of the Paul Sabatier University, Toulouse, France. The problem is first reduced to an unconstrained one using penalty techniques. The objective function thus has its domain partitioned into cells over each of which it is continuous. Conditions characterizing a minimum are developed. In the first-order algorithm, descent directions are obtained by using an active set

approach. Univariate minimization is achieved via a specialized line search which recognizes the possibility of a first derivative discontinuity or jump in the function. The second order algorithm deals with the projection of the Newton direction. The numerical linear algebra techniques involve the computation of orthogonal and nonorthogonal projectors, and their updates. A perturbation method is used to handle degeneracy and the approach is shown to be practical.

Applications include VLSI, nonlinear robust regression and operations research/industrial engineering.

This is a difficult area of some importance for which there is much need for applicable algorithms. Moreover, the area bridges certain aspects of both continuous and discrete optimization.

We are currently looking at a fixed charge problem applied to assembly line production, originally considered by Hiraki in the Japanese OR journal. We have also been successful at tackling a mixed problem involving knapsack constraints, 0-1 variables representing allocations and a quadratic model for the generation of power. This latter is an application originally investigated by Eric Klein of Ontario Hydro, Canada and is to appear in EJOR. Dr. Mongeau is maintaining contact with Dr. Klein and is looking at more general problems. He is also working on a non-separable fixed charge application to transportation and a cutting stock problem of Haberl, that was published in JOTA in 1991 and has recently been generalized in a submission to Mathematical Programming.

Our algorithm works in a significantly smaller dimensional space than the approach of Hiraki, and consequently we expect to be able to tackle much larger problems than he is able to handle. Furthermore, our approach is more general, since Hiraki's method requires assumptions that are rather artificial.

These results appear in Conn and Mongeau (1993a) and Conn and Mongeau (1993b). A synthesis of these two papers, but restricted to the linear case is to be published in Conn and Mongeau (1995). The application to discharge allocation for hydro-electric generating stations is most interesting in that we are able to consider a more realistic nonlinear model rather than the usual linear model. In particular, it is known that in the operating domain (where the discharge is sufficiently large) the quantity of electricity generated by one particular generating unit is a quadratic polynomial function of the water drafted through the unit.

5 Technical Publications

Included below is a summary of the twenty technical publications to date that have benefitted from and acknowledge the AKPA support. I will send copies of the full reports under separate cover.

1) A.R. Conn, N.I.M. Gould and Ph.L. Toint, Large-Scale Nonlinear Constrained Optimization

During the past ten years, much progress has been made in the theory and practice of constrained nonlinear optimization. However, considerable obstacles appear when these ideas are applied to large-scale problems. This is important as many real applications require the solution of problems in thousands of unknowns. In some areas, in particular linear programming, considerable progress has been made. But even modest departures into nonlinearity, for example the solution of large, general quadratic programs, present considerable challenges. This is apparent

when one views the paucity of software for solving such problems. Unsurprisingly, the position does not improve as more drastic forms of nonlinearity are encountered.

In this paper, we will try to explain why the difficulties arise, how attempts are being made to overcome them and what the problems are that still remain.

(Published in ICIAM '91: Proceedings of the second international conference on industrial and applied mathematics, Washington, Ed R.E. O'Malley, Jr. 1992 pp. 51-70, SIAM)

2] A.R. Conn, N.I.M. Gould and Ph.L. Toint, Numerical experiments with the LANCELOT package (Release A) for large-scale nonlinear optimization

In this paper, we describe the algorithmic options of Release A of LANCELOT, a new Fortran package for large-scale nonlinear optimization. We then present the results of intensive numerical tests and discuss the relative merits of the options. The experiments described involve both academic and applied problems. Conclusions are finally proposed, both specific to LANCELOT and of more general scope.

(To appear in Mathematical Programming)

3] A.R. Conn, N.I.M. Gould and Ph.L. Toint, Complete numerical results for large-scale tests on LANCELOT Release A

This report contains the detailed results of the numerical experiments on the LANCELOT package for nonlinear optimization (Release A). These results constitute the basis for the discussion and analysis presented in the report [2] above.

4] I. Bongartz, A.R. Conn, N.I.M. Gould and Ph.L. Toint, CUTE : Constrained and unconstrained testing environment

The purpose of this paper is to discuss the scope and functionality of a versatile environment for testing small and large-scale nonlinear optimization algorithms. Although many of these facilities were originally produced by the authors in conjunction with the software package LANCELOT, we believe that they will be useful in their own right and should be available to researchers for their development of optimization software. The tools are available by anonymous ftp from a number of sources and may, in many cases, be installed automatically.

The scope of a major collection of test problems written in the standard input format (SIF) used by the LANCELOT software package is described. Recognizing that most software was not written with the SIF in mind, we provide tools to assist in building an interface between this input format and other optimization packages. These tools already provide a link between the SIF and an number of existing packages, including MINOS and OSL. In addition, as each problem includes a specific classification that is designed to be useful in identifying particular classes of problems, facilities are provided to build and manage a database of this information.

(Published in ACM Transactions on Mathematical Software, Volume 21, pp123-160, 1995)

5] A.R. Conn, N.I.M. Gould and Ph.L. Toint, A note on exploiting structure when using slack variables

We show how to exploit the structure inherent in the linear algebra for constrained nonlinear optimization problems when inequality constraints have been converted to equations by adding slack variables.

(Published in Mathematical Programming, Volume 67, pp89-97, 1994)

6] I. Bongartz, P.H. Calamai and A.R. Conn A projection method for l_p norm location-allocation problems

We present a solution method for location-allocation problems involving the l_p norm, where $1 < p < \infty$. The method relaxes the 0,1 constraints on the allocations, and solves for both the locations and allocations simultaneously. Necessary and sufficient conditions for local minima of the relaxed problem are stated and used to develop an iterative algorithm. This algorithm finds a stationary point on a series of subspaces defined by the current set of activities. The descent direction is a projection onto the current subspace of a direction incorporating second-order information for the locations, and first-order information for the allocations. Under mild conditions, the algorithm finds local minima with 0,1 allocations and exhibits quadratic convergence. An implementation that exploits the special structure of these problems to dramatically reduce the computational cost of the required numerical linear algebra is described. Numerical results for thirty-six test problems are included.

(Published in Mathematical Programming, Volume 66, pp283-312, 1994)

7] A.R. Conn, N.I.M. Gould, A. Sartenaer and Ph.L. Toint Global Convergence of two Augmented Lagrangian Algorithms for Optimization with a Combination of General Equality and Linear Constraints

We consider the global convergence properties of a class of augmented Lagrangian methods for solving nonlinear programming problems. In the proposed method, linear constraints are treated separately from more general constraints. Thus only the latter are combined with the objective function in an augmented Lagrangian. The subproblem then consists of (approximately) minimizing this augmented Lagrangian subject to the linear constraints. In this paper, we prove the global convergence of the sequence of iterates generated by this technique to a first-order stationary point of the original problem. We consider various stopping rules for the iterative solution of the subproblem, including practical tests used in several existing packages for linearly constrained optimization. We also extend our results to the case where the augmented Lagrangian's definition involves several distinct penalty parameters.

8] A.R. Conn, N.I.M. Gould, A. Sartenaer and Ph.L. Toint Local Convergence Properties of two Augmented Lagrangian Algorithms for Optimization with a Combination of General Equality and Linear Constraints

We consider the local convergence properties of the class of augmented Lagrangian methods for solving nonlinear programming problems whose global convergence properties are analyzed by A.R. Conn, N.I.M. Gould, A. Sartenaer and Ph.L. Toint in "Global Convergence of two Augmented Lagrangian Algorithms for Optimization with a Combination of General Equality and Linear Constraints" In these methods, linear constraints are treated separately from more general constraints. These latter constraints are combined with the objective function in an augmented Lagrangian while the subproblem then consists in (approximately) minimizing this augmented Lagrangian subject to the linear constraints. The stopping rule that we consider for this inner iteration covers practical tests used in several existing packages for linearly constrained optimization. Our algorithmic class also allows several distinct penalty parameters associated with different subsets of general equality constraints. In this paper, we analyze the local convergence

of the sequence of iterates generated by this technique and prove fast linear convergence and boundedness of the potentially troublesome penalty parameters.

9] A.R. Conn, N.I.M. Gould, A. Sartenaer and Ph.L. Toint Convergence Properties of an Augmented Lagrangian Algorithms for Optimization with a Combination of General Equality and Linear Constraints

We consider the global and local convergence properties of a class of augmented Lagrangian methods for solving nonlinear programming problems. In these methods, linear and more general constraints are handled in different ways. The general constraints are combined with the objective function in an augmented Lagrangian. The iteration consists of solving a sequence of subproblems; in each subproblem the augmented Lagrangian is approximately minimized in the region defined by the linear constraints. A subproblem is terminated as soon as a stopping condition is satisfied. The stopping rules that we consider here encompass practical tests used in several existing packages for linearly constrained optimization. Our algorithm also allows different penalty parameters to be associated with disjoint subsets of the general constraints. In this paper, we analyze the convergence of the sequence of iterates generated by such an algorithm and prove global and fast linear convergence as well as showing that potentially troublesome penalty parameters remain bounded away from zero.

(To appear in SIAM J. on Optimization)

10] A.R. Conn and M. Mongeau, Discontinuous Piecewise Differentiable Optimization I: Theory

A theoretical framework and a practical algorithm are presented to solve discontinuous piecewise linear optimization problems. A penalty approach allows one to consider such problems subject to a wide range of constraints involving piecewise linear functions. Although the theory is expounded in detail in the special case of discontinuous piecewise linear functions, it is straightforwardly extendable, using standard nonlinear programming techniques, to nonlinear (discontinuous piecewise differentiable) functions.

This work is presented in two parts. We introduce the theory in this first paper. The descent algorithm which is elaborated uses active-set and projected gradient approaches. It is a generalization of the ideas used by Conn to deal with nonsmoothness in the l_1 exact penalty function, and it is based on the notion of decomposition of a function into a smooth and a nonsmooth part.

In an accompanying paper, we shall tackle constraints via a penalty approach, we shall discuss the degenerate situation, and numerical results will be presented.

11] A.R. Conn and M. Mongeau, Discontinuous Piecewise Differentiable Optimization II: Degeneracy and Applications

The implementation of an algorithm, developed in "Discontinuous Piecewise Differentiable Optimization I: Theory" by A.R. Conn and M. Mongeau, to solve discontinuous piecewise linear optimization problems is discussed, and limited numerical applications are presented.

A penalty approach allows one to consider such problems subject to a wide range of constraints involving piecewise linear functions. The constrained case is reduced to the unconstrained minimization of a (piecewise linear) l_1 exact penalty function. We also discuss how the algorithm is

modified when it encounters degenerate points and other points where the objective function is not decomposable. The algorithm is applied to discontinuous optimization problems from models in industrial engineering and manufacturing.

12] A.R. Conn and M. Mongeau, Discontinuous Piecewise Linear Optimization

A theoretical framework and a practical algorithm are presented to solve discontinuous piecewise linear optimization problems dealing with functions for which the *ridges* are known. A penalty approach allows one to consider such problems subject to a wide range of constraints involving piecewise linear functions. Although the theory is expounded in detail in the special case of discontinuous piecewise *linear* functions, it is straightforwardly extendable, using standard nonlinear programming techniques, to *nonlinear* (discontinuous piecewise differentiable) functions.

The descent algorithm which is elaborated uses active-set and projected gradient approaches. It is a generalization of my earlier work to deal with nonsmoothness in the l_1 exact penalty function, and it is based on the notion of *decomposition* of a function into a smooth and a nonsmooth part. A penalty approach allows one to consider such problems subject to a wide range of constraints involving piecewise linear functions. The constrained case is reduced to the unconstrained minimization of a (piecewise linear) l_1 exact penalty function. We also discuss how the algorithm is modified when it encounters degenerate points. Preliminary numerical results are presented: the algorithm is applied to discontinuous optimization problems from models in manufacturing.

(Submitted to Mathematical Programming)

13] A.R. Conn, N.I.M. Gould and Ph.L. Toint, Improving the decomposition of partially separable functions in the context of large-scale optimization: a first approach.

This paper examines the question of modifying the decomposition of a partially separable function in order to improve computational efficiency of large-scale minimization algorithms using a conjugate-gradient inner iteration. The context and motivation are given and the application of a simple strategy discussed on examples extracted from the CUTE test problem collection.

(Published in "Large-Scale Optimization: State of the Art", W.W. Hager, D.W. Hearn and P.M. Pardalos, eds., Kluwer Academic Publishers, pp82-94, 1994)

14] A.R. Conn, N.I.M. Gould, A. Sartenaer and Ph. Toint On Iterated-Subspace Minimization Methods for Nonlinear Optimization

This paper examines a class of Iterated-Subspace Minimization (ISM) methods for solving large-scale unconstrained minimization problems. At each major iteration of such a method, a low-dimensional manifold, the iterated subspace, is constructed and an approximate minimizer of the objective function in this manifold then determined. The iterated subspace is chosen to contain vectors which ensure global convergence of the overall scheme and may also contain conjugate gradient related vectors which encourage fast asymptotic convergence. We demonstrate that this approach can sometimes be very advantageous and indicate the general performance on a collection of large problems. Moreover, comparisons with a limited memory approach and LANCELOT are made.

(To appear in "Proceedings on Linear and Nonlinear Conjugate Gradient-Related Methods",

L. Adams and L. Nazareth, eds., SIAM, 1996)

15] A.R. Conn, N.I.M. Gould and Ph.L. Toint, Algorithms for large-scale constrained nonlinear optimization : a current survey

Much progress has been made in constrained nonlinear optimization in the past ten years, but most large-scale problems still represent a considerable obstacle.

In this survey paper we attempt to give an overview of the current approaches, including interior and exterior methods and algorithms based upon trust regions and line searches. In addition, the importance of software, numerical linear algebra and testing is addressed. We try and explain why the difficulties arise, how attempts are being made to overcome them and some of the problems that still remain.

Although there is some emphasis on the LANCELOT and CUTE projects, the intention is to give a broad picture of the state-of-the-art.

(Published in "Algorithms for continuous optimization:the state of the art", E. Spedicato, ed., pp287-332, Volume 434 of NATO ASI Series C: Mathematical and Physical Sciences, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994)

16] A.R. Conn and Ph.L. Toint, An algorithm using quadratic interpolation for unconstrained derivative free optimization

This paper explores the use of multivariate interpolation techniques in the context of methods for unconstrained optimization that do not require derivative of the objective function. A new algorithm is proposed that uses quadratic models in a trust region framework. The algorithm is constructed to require few evaluations of the objective function and is designed to be relatively insensitive to noise in the objective function values. Its performance is analyzed on a set of twenty examples, both with and without noise.

(To appear in "Nonlinear Optimization and Applications"; G. Di Pillo and F. Giannessi, eds., Plenum Publishing, 1995)

17] A.R. Conn, N.I.M. Gould and Ph.L. Toint, A globally convergent Lagrangian barrier algorithm for optimization with general inequality constraints and simple bounds

In this manuscript we consider the global and local convergence properties of a class of Lagrangian barrier methods for solving nonlinear programming problems. In such methods, simple bound constraints may be treated separately from more general constraints. The objective and general constraint functions are combined in a Lagrangian barrier function. A sequence of such functions are approximately minimized within the domain defined by the simple bounds. Global convergence of the sequence of generated iterates to a first-order stationary point for the original problem is established. Furthermore, possible numerical difficulties associated with barrier function methods are avoided as it is shown that a potentially troublesome penalty parameter is bounded away from zero. This paper is a companion to our previous work (see, Conn *et al.*, 1991) on augmented Lagrangian methods.

(To appear in Mathematics of Computation, 1995)

18] B. Reiser and B. J. Flehinger and A. R. Conn, Estimating component defect probability from masked system success/failure data

We consider the situation where a system consisting of k components will fail whenever there

is a defect in one or more of the components. Due to cost and time constraints it may not always be feasible to learn exactly which components are defective. Instead, test procedures may ascertain that the defective components belong to some subset of the k components. This phenomenon is termed masking. Given historical masked data we would like to estimate the quality or defect probability of each individual component. In addition, when masked data will occur in the future, we would like to diagnose the defective components. In this manuscript we discuss the analysis of such data and suggest a two-stage experimental procedure which permits the calculation of maximum likelihood estimates of the probabilities of interest.

(Submitted to IEEE Transactions on Reliability)

19] A.R. Conn, N.I.M. Gould and Ph.L. Toint, A note on using alternative second-order models for the subproblems arising in barrier function methods for minimization

Inequality constrained minimization problems are often solved by considering a sequence of parameterized barrier functions. Each barrier function is approximately minimized and the relevant parameters subsequently adjusted. It is common for the estimated solution to one barrier function problem to be used as a starting estimate for the next. However, this has unfortunate repercussions for the standard Newton-like methods applied to the barrier subproblem. In this paper, we consider a class of alternative Newton methods which attempt to avoid such difficulties. Such schemes have already proved of use in the Harwell Subroutine Library quadratic programming codes VE14 and VE19.

(Published in Numerische Mathematik, Volume 68, pp17-33, 1994)

20] A.R. Conn, N.I.M. Gould, A. Sartenaer and Ph.L. Toint, Convergence properties of minimization algorithms for convex constraints using a structured trust region

In this paper, we present a class of trust region algorithms for minimization problems within convex feasible regions, in which the structure of the problem is explicitly used in the definition of the trust region. This development is intended to reflect the possibility that some parts of the problem may be more accurately modeled than others, a common occurrence in large-scale nonlinear applications. After describing the structured trust region mechanism, we prove global convergence for all algorithms in our class.

(To appear in SIAM J. on Optimization, 1995)

6 FINANCIAL

The final financial statement will be sent under separate cover.

6.1 EPILOGUE

Our general objectives in this research have been twofold. Our primary objective has always been to apply optimization to real problems in manufacturing. Our secondary objective is to thereby advance research into nonlinear programming algorithms. In order to achieve the first objective we consider it essential to work directly with the personnel in the manufacturing environment. We actively seek out manufacturing groups and we request from them the problems which they consider important. Thus the framework in which we work consists of initially famil-

iarizing ourselves with a particular production team, followed by identifying suitable problems, in consideration of both the applicability of our expertise and the significance of the problem in the manufacturing environment. The next stage consists of producing an initial model that is simple enough to give rapid feedback (in order to maintain and stimulate cooperation) and realistic enough to indicate useful results (in order to encourage confidence in the approach). This stage also includes the collection of data from the practitioners. Once the initial model has been solved, one goes through a cycle of analyzing results, expanding and modifying the model, re-solving, collecting new data, etc. The ultimate goal is to see the solution implemented in the operation of the manufacturing site. During the period of this award we have been able to initiate a considerable number of projects exactly within the model outlined above.

Given the current progress we have no doubt that our work in nonlinear programming can make useful contributions to manufacturing and we intend to do our utmost to pursue these approaches. In addition, we have achieved extremely useful input and support for our more theoretical endeavors.

6.2 APPENDIX

During the tenure of this grant my colleagues N. I. M. Gould, Ph. L. Toint and myself were awarded the Beale/Orchard-Hays Prize (the foremost computational award in mathematical programming). Since a substantial amount of the fundamental research carried out under the auspices of this grant involves LANCELOT and CUTE I decided to include the entire citation in an appendix.

Beale/Orchard-Hays Prize official nomination

Presented at the 15th International Mathematical Programming Symposium, Ann Arbor
(USA), 15th August 1994

After a very close competition involving many excellent nominations, the Beale/Orchard-Hays prize committee is pleased to announce the award of the 1994 prize to Andrew R. Conn of I.B.M., Yorktown Heights, Nicholas I.M. Gould of the Appleton Rutherford Laboratory, Oxfordshire and Philippe L. Toint of the Facultes Universitaires Notre-Dame de la Paix, Namur for their book entitled:

LANCELOT: A Fortran Package for Large-Scale Nonlinear Optimization, Springer Verlag, Berlin 1992.

As the authors say in their introduction, LANCELOT (Large And Nonlinear Constrained Extended Lagrangian Optimization Techniques) was created out of the necessity for accurate modeling of physical, scientific, statistical and economic phenomena, which led to ever larger and more challenging nonlinear optimization problems. However, necessity needs to be recognized, and complemented by the courage, determination, and ability to undertake a major research effort spread over several years and nations.

To summarize briefly several of the achievements in the course of this project:

1. significant work on the structure of large scale nonlinear optimization problems, and on algorithmic approaches for large problems i.e. their joint paper "A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds", SIAM Journal on Numerical Analysis 28, 545-572 (1991).
2. the development of a uniform input representation for nonlinear programs. The authors were not enamored of the MPS format, but were convinced by the user community that if LANCELOT were to get the widest possible use on real problems, total compatibility with the mps format was required. Thus they deserve commendation.
3. the development of a code sufficiently robust that it efficiently solves without tuning the widest range of nonlinear programming problems of any code yet devised.
4. the creation and distribution of the CUTE test environment, which is a major asset to the field.
5. the quality of the documentation provided by the book.
6. the decision to distribute LANCELOT in return for a new problem from the user, thus making the software easily available and simultaneously promoting research.

In conclusion we hope and believe that LANCELOT is but a beginning, and given that necessity has received such a helping hand, many others in the field will be encouraged to tackle the many problems crying out for a solution.

Robert R. Meyer

David F. Shanno

Robert Vanderbei

Laurence A. Wolsey (Chair)

References

- [Anderson and Christiansen, 1995] K.D. Anderson and E. Christiansen. A Newton barrier method for minimizing a sum of Euclidean norms subject to linear equality constraints. Technical report, Departement of Mathematics and Computer Science, Odense University, Odense, Denmark, 1995.
- [Bongartz et al., 1994] I. Bongartz, P. H. Calamai, and A. R. Conn. A projection method for ℓ_p norm location-allocation problems. *Mathematical Programming*, 66(3):283-312, 1994.
- [Bongartz et al., 1995a] I. Bongartz, A. R. Conn, N. I. M. Gould, M. Saunders, and Ph. L. Toint. A numerical comparison between the lancetot and MINOS packages for large-scale nonlinear optimization. Research Report, (in preparation), IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA, 1995.

- [Bongartz *et al.*, 1995b] I. Bongartz, A. R. Conn, N. I. M. Gould, and Ph. L. Toint. CUTE: Constrained and unconstrained testing environment. *ACM Transactions on Mathematical Software*, 21(1):123 – 160, 1995.
- [Chan *et al.*, 1995] T. F. Chan, G. H. Golub, and P. Mulet. A nonlinear primal-dual method for tv based image reconstruction. Technical report, Computational and Applied Mathematics, Department of Mathematics, UCLA, Los Angeles, CA 90024-1555, USA, 1995.
- [Conn and Mongeau, 1993a] A. R. Conn and M. Mongeau. Discontinuous piecewise differentiable optimization I: Theory. Technical Report CRM-1868, Centre de recherches mathématiques, Université de Montréal, CP 6128-A, H3C 3J7 Canada, 1993.
- [Conn and Mongeau, 1993b] A. R. Conn and M. Mongeau. Discontinuous piecewise differentiable optimization II: Degeneracy and applications. Technical Report CRM-1869, Centre de recherches mathématiques, Université de Montréal, CP 6128-A, H3C 3J7 Canada, 1993.
- [Conn and Mongeau, 1995] A. R. Conn and M. Mongeau. Discontinuous piecewise linear optimization. *Mathematical Programming*, (to appear), 1995.
- [Conn and Toint, 1994] A. R. Conn and Ph. L. Toint. A comparison for small-scale nonlinear optimization software (CSSNOS): Preliminary project outline. Technical Report 94/4, Department of Mathematics, FUNDP, Namur, Belgium, 1994.
- [Conn and Toint, 1995] A. R. Conn and Ph. L. Toint. An algorithm using quadratic interpolation for unconstrained derivative free optimization. In Gianni Di Pillo and Franco Giannessi, editors, *Nonlinear Optimization and Applications*. Plenum Publishing, (to appear), 1995.
- [Conn *et al.*, 1988] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. Testing a class of methods for solving minimization problems with simple bounds on the variables. *Mathematics of Computation*, 50:399–430, 1988.
- [Conn *et al.*, 1991] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Numerical Analysis*, 28(2):545–572, 1991.
- [Conn *et al.*, 1993a] A. R. Conn, Nick Gould, A. Sartenaer, and Ph. L. Toint. Global convergence of two augmented Lagrangian algorithms for optimization with a combination of general equality and linear constraints. Research Report RC18900, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA, 1993.
- [Conn *et al.*, 1993b] A. R. Conn, Nick Gould, A. Sartenaer, and Ph. L. Toint. Local convergence properties of two augmented Lagrangian algorithms for optimization with a combination of general equality and linear constraints. Research Report RC18901, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA, 1993.
- [Conn *et al.*, 1994a] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. Improving the decomposition of partially separable functions in the context of large-scale optimization: a first approach. In

- W. W. Hager, D. W. Hearn, and P.M. Pardalos, editors, *Large Scale Optimization: State of the Art*, pages 82-94. Kluwer Academic Publishers, 1994.
- [Conn et al., 1994b] A. R. Conn, Nick Gould, and Ph. L. Toint. A note on exploiting structure when using slack variables. *Mathematical Programming*, 67(1):89-97, 1994.
- [Conn et al., 1994c] A. R. Conn, Nick Gould, and Ph. L. Toint. A note on using alternative second-order models for the subproblems arising in barrier function methods for minimization. *Numerische Mathematik*, 68(1):17-33, 1994.
- [Conn et al., 1995a] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. A globally convergent Lagrangian barrier algorithm for optimization with general inequality constraints and simple bounds. *Mathematics of Computation*, (to appear), 1995.
- [Conn et al., 1995b] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. Numerical experiments with the lancet package (release a) for large-scale nonlinear optimization. *Mathematical Programming*, (to appear), 1995.
- [Conn et al., 1995c] A. R. Conn, Nick Gould, A. Sartenaer, and Ph. L. Toint. Convergence properties of an augmented Lagrangian algorithms for optimization with a combination of general equality and linear constraints. *SIAM Journal on Optimization*, (to appear), 1995.
- [Conn et al., 1995d] A. R. Conn, Nick Gould, A. Sartenaer, and Ph. L. Toint. Convergence properties of minimization algorithm for convex constraints using a structured trust region. *SIAM Journal on Optimization*, (to appear), 1995.
- [Conn et al., 1995e] A. R. Conn, Nick Gould, A. Sartenaer, and Ph. L. Toint. A globally convergent Lagrangian barrier algorithm for optimization with a combination of general inequality and linear constraints. Technical Report, (in preparation), Department of Mathematics, FUNDP, Namur, Belgium, 1995.
- [Conn et al., 1995f] A. R. Conn, Nick Gould, A. Sartenaer, and Ph. L. Toint. Local convergence properties of a Lagrangian barrier algorithm for optimization with a combination of general inequality and linear constraints. Research Report, (in preparation), IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA, 1995.
- [Dennis and Schnabel, 1983] J. E. Dennis and R. B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. Prentice-Hall, Englewood Cliffs, NJ, 1983.
- [Efron and Tibshirani, 1986] B. Efron and R. Tibshirani. Bootstrap methods for statistical errors, confidence intervals and other measures of statistical accuracy. *Statistical Science*, 1(1):54-74, 1986.
- [Mates, 1993] D. M. Mates. *A projection method for the floor planning problem*. PhD thesis, Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada, 1993.

[Nelder and Mead, 1965] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308-313, 1965.

[Powell, 1994a] M. J. D. Powell. A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in optimization and numerical analysis, Proceedings of the Sixth workshop on Optimization and Numerical Analysis, Oaxaca, Mexico*, volume 275 of *Mathematics and its Applications*, pages 51-67. Kluwer Academic Publishers, 1994.

[Powell, 1994b] M. J. D. Powell. A direct search optimization method that models the objective by quadratic interpolation. Presentation at the 5th Stockholm Optimization Days, 1994

[Reiser *et al.*, 1994] B. Reiser, B. J. Flehinger, and A. R. Conn. Estimating component defect probability from masked system success/failure data. Research Report RC 19720, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA, 1994.

[The Mathworks Inc, 1992] The Mathworks Inc. *Matlab reference guide*. The Mathworks Inc., Natick, Mass. USA, 1992.